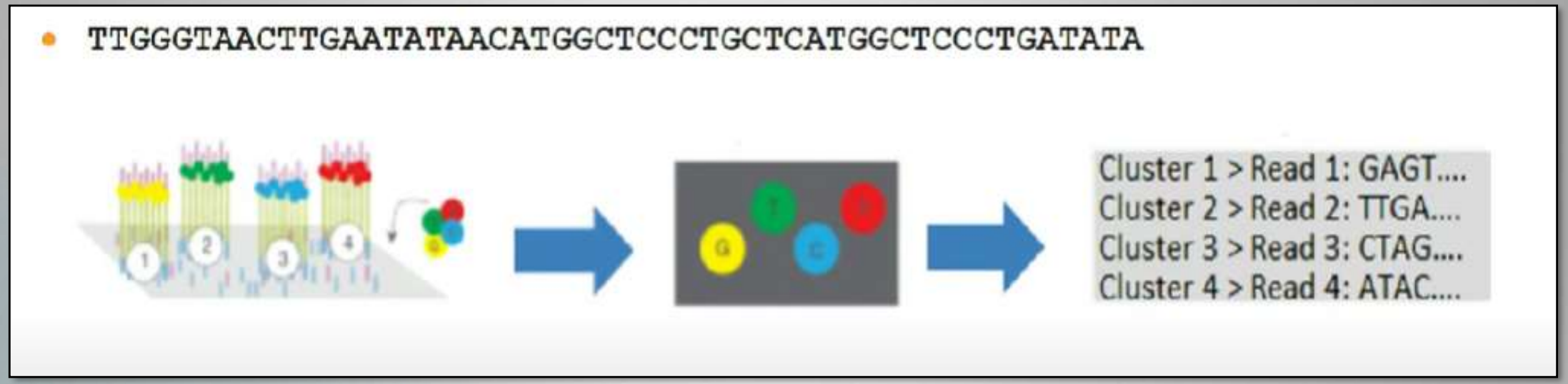
Types of Reads

- Single Read
- Pair End Read
- Mate Pair Read

What is Read?

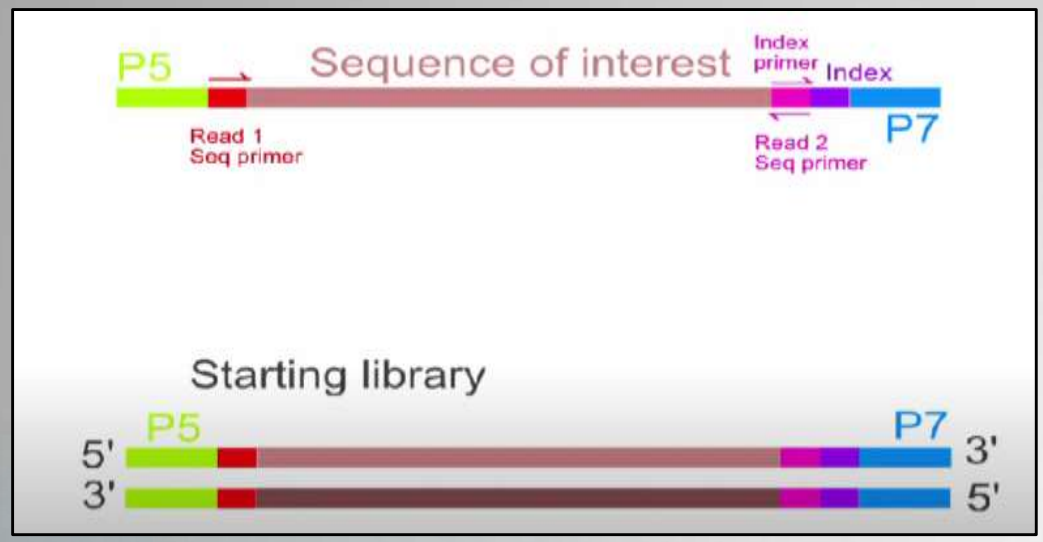
- Read is the data that comes from a single cluster on the flow cell.

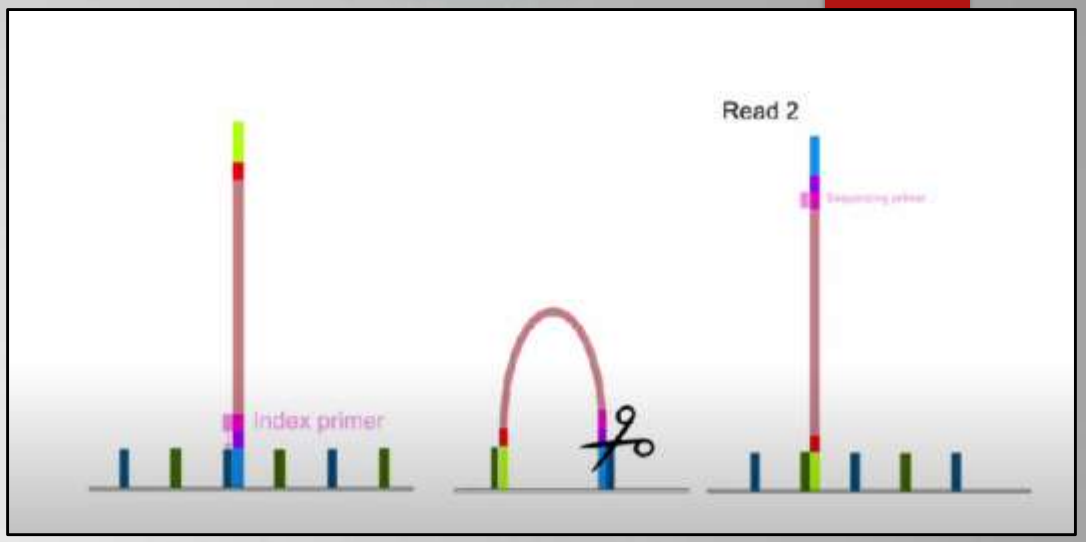


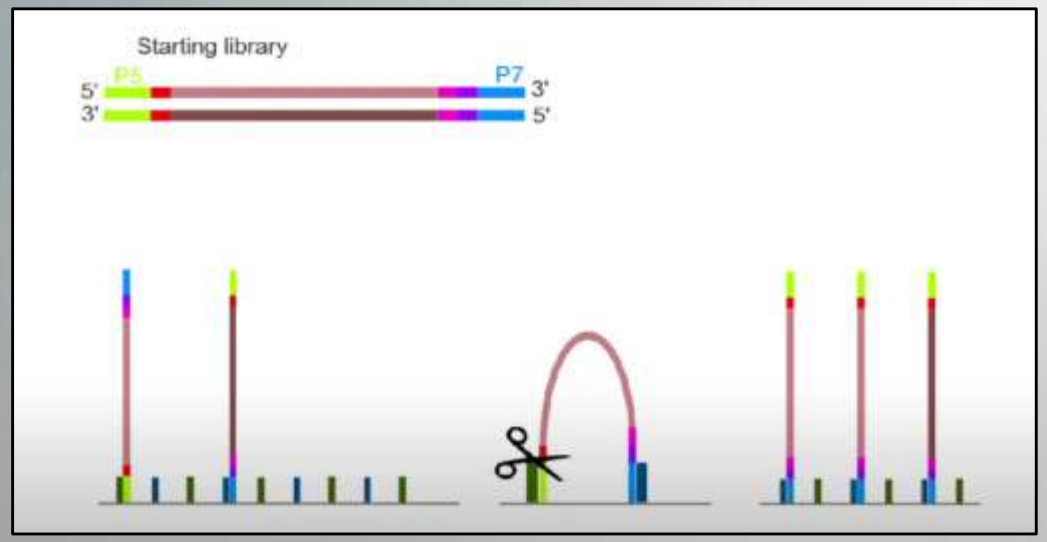
Reads are generally stored in FASTQ files.

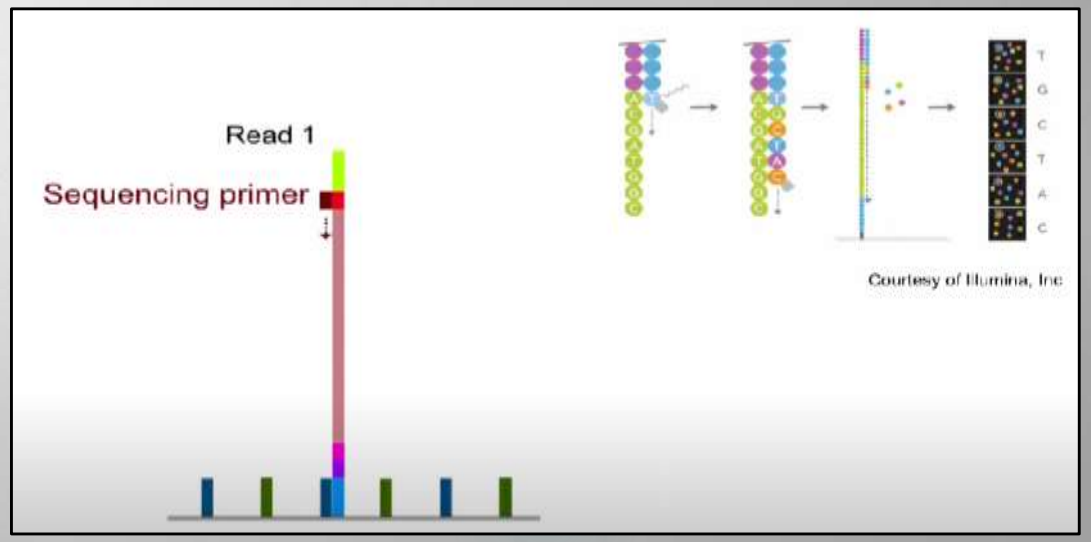
- Input for analysis pipelines
- Demultiplexing

How does single read and pair end sequencing done?







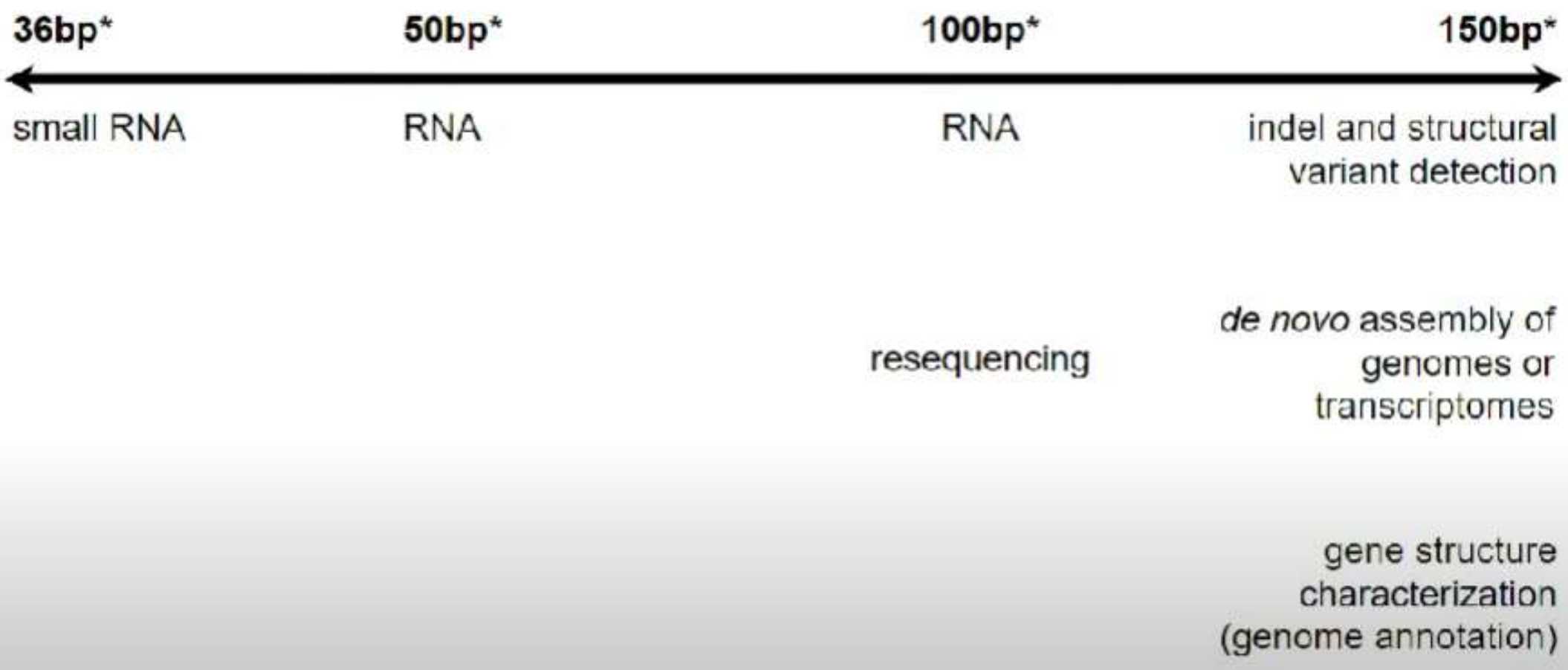


Single-end sequencing	Paired-end sequencing
Sequence from the single end.	Sequence from both ends.
From 5' to 3' direction.	From 5' to 3' and 3' to 5' direction.
A single set of flow cell complementary oligos, index, and adaptor sequence is used.	A pair of flow cell complementary oligos, index, and adaptor sequence is used on either end.
A single read is generated.	Two reads are generated.
Known as a forward read or R1.	Known as forward and reverse reads or R1 and R2.
A single FASTQ file is generated which is FASTQR.	Two different FASTQ files are generated which is FASTQR1 and FASTQR2.
Not suitable for de novo sequencing.	Suitable for de novo sequencing.
Not suitable to sequence larger repetitive regions.	Suitable to sequence larger repetitive regions.
Suitable for RNA-seq and ChIP-seq.	Suitable for both DNA and RNA sequencing.
Cheper.	Costlier.

Will single reads be sufficient or are paired end reads required?

Application	SR or PE?	Notes
SNP Detection (Resequencing)	Either	Coverage depth is key
Indel or Structure Variant Detection (Resequencing)	PE	Analysis methods are based on PE data
De Novo Genome or Transcriptome Assembly	PE	PE info is used in assembly process
RNA-Seq (Expression)	Either	PE needed for identification of novel transcripts and gene structure characterization
Small RNA Differential Expression	SR	PE will result in high overlap

What is Read Length?



FASTQ Files - Overview

Single Index read

Dual Index read

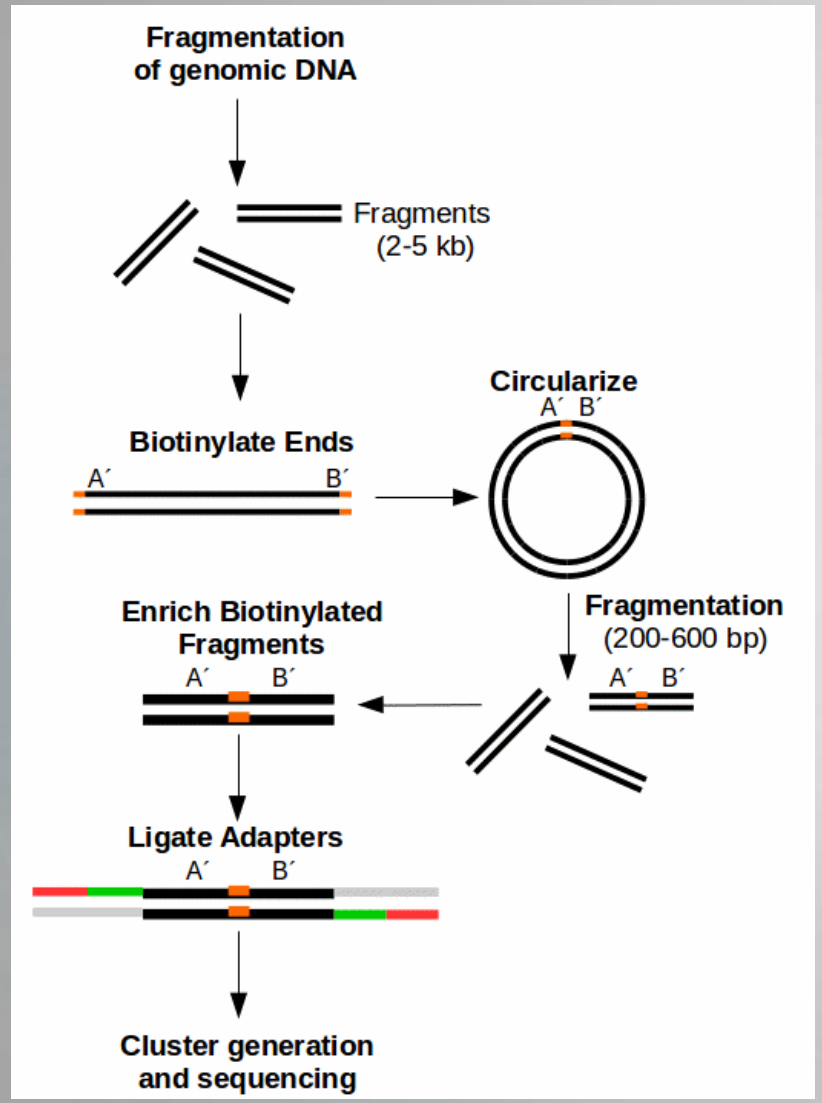
Ex.

```
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13501:2240 1:N:0 CTTGTA
TGAAACCAGTGTTCTTAATTGGCATTTTACACACACACACACAGAATTTAAAAAAAAAATCAAAGG
+
=55>7;?::BDADDD@EE88DCD?DFFEFFECBE6666EB=B;<;<-34:;<CB51>=BEEE>EE?
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13660:2247 1:N:0 CTTGTA
CCAAACATTAAGTAACTCTTAAAATGGCACACAGGTTTTAAAGCTATTGGTTTTTCCTTCCTAACT
+
FFEDFBGEGGGGDFGEFFFFGGDF=FBFFFGGGE7CEEDEFBFBFGEEGF@FCDDFDFFEGFEAGF
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13966:2183 1:N:0 CTTGTA
TTGGGTAACTTGAATATAACATGGCTCCCTTGCTGTAAGCAAATGTTTTAGAGCTGAATTTTTCCT
+
HHHHHEHHHHHHFHHHHHHHHHHHHHHHHHHHHGGFHHHHHHHHHHFHHHFHEHHFHEHHHHFHHHF
```

```
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13501:2240 1:N:0 CTTGTA+TGAATA
TGAAACCAGTGTTCTTAATTGGCATTTTACACACACACACACAGAATTTAAAAAAAAAATCAAAGG
+
=55>7;?::BDADDD@EE88DCD?DFFEFFECBE6666EB=B;<;<-34:;<CB51>=BEEE>EE?
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13660:2247 1:N:0 CTTGTA+TGAATA
CCAAACATTAAGTAACTCTTAAAATGGCACACAGGTTTTAAAGCTATTGGTTTTTCCTTCCTAACT
+
FFEDFBGEGGGGDFGEFFFFGGDF=FBFFFGGGE7CEEDEFBFBFGEEGF@FCDDFDFFEGFEAGF
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13966:2183 1:N:0 CTTGTA+TGAATA
TTGGGTAACTTGAATATAACATGGCTCCCTTGCTGTAAGCAAATGTTTTAGAGCTGAATTTTTCCT
+
HHHHHEHHHHHHFHHHHHHHHHHHHHHHHHHHHGGFHHHHHHHHHHFHHHFHEHHFHEHHHHFHHHF
```

- Reads are generally stored in FASTQ files
- Each file contains sequences and quality scores
- Each file can contain millions of reads
- Number of reads x Read length = Number of base pairs

Mate Pair sequencing



- Allows us to obtain pair end reads with long inserts.
- The produced fragment(200-600bp) contains the end of the original long fragment and can be sequenced.
- After sequencing, we'll get information about original fragment.

Applications:

- Genome assembly
- Structural variant detection
- Gene Fusion Identification
- Repeat analysis
- Genome structure variation profiling

Example of Mate Pair read

```
>Read1
Sequence: GATCAGTACGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
Quality Scores: !""#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNO
>Read2
Sequence: CGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
Quality Scores: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
```

- Read1 and Read2 represent the paired-end reads obtained from mate pair sequencing.
- Each read consists of a nucleotide sequence (Sequence) and its corresponding quality scores.
- The nucleotide sequence represents the actual DNA bases obtained from sequencing, while the quality scores represent the confidence or accuracy of each base call.

References

- https://www.youtube.com/watch?v=XdoAnsDPMfA&ab_channel=Illumina
- https://www.youtube.com/watch?v=HMyCqWhwB8E&t=10s&ab_channel=Illumina
- https://www.youtube.com/watch?v=9ezaTbOVHYQ&ab_channel=BMHlearning
- https://emea.illumina.com/science/technology/next-generation-sequencing/mate-pair-sequencing.html
- https://www.novogene.com/eu-en/resources/blog/sequencing-by-synthesis-on-the-illumina-novaseqxtm-series/

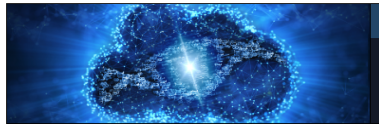
DATABASES OF NGS

- National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).
- European Bioinformatics Institute (EBI) European Nucleotide Archive(ENA).

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA | SRA | Search
Advanced | Help



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started
- Documentation
- How to submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software
- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources
- Submission Portal
- dbGaP Home
- BioProject
- BioSample

FOLLOW NCBI

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA | SRA | Bioreactor | Search
Create alert | Advanced | Help

Access
Public (35,222)

Source
DNA (32,157)
RNA (2,398)

Type
exome (26)
genome (3,084)

Library Layout
paired (28,646)
single (6,623)

Platform
ABI SOLiD (2)
BGISEQ (40)
Capillary (2)
Illumina (30,286)
Ion Torrent (1,725)
LS454 (2,901)
Oxford Nanopore (159)
PacBio SMRT (137)

Strategy
Exome (547)
Genome (3,389)
RNASeq (11)
other (31,322)

Data in Cloud
GS (35,208)
S3 (35,209)

File Type
bam (307)
fastq (28,821)
sff (541)

Other
aligned data (52)

Clear all

Summary | 20 per page | Send to:

Filters: Manage Filters

Search results

Items: 1 to 20 of 35269

<< First | < Prev | Page 1 of 1764 | Next > | Last >>

1. salinity and heavy metal sludge
   1 ILLUMINA (Illumina NovaSeq 6000) run: 35.8M spots, 10.8G bases, 3.4Gb downloads
   Accession: SRX24151668
2. heavy metal sludge
   1 ILLUMINA (Illumina NovaSeq 6000) run: 34.5M spots, 10.4G bases, 3.3Gb downloads
   Accession: SRX24151667
3. salinity sludge
   1 ILLUMINA (Illumina NovaSeq 6000) run: 35.1M spots, 10.6G bases, 3.3Gb downloads
   Accession: SRX24151666
4. Control sludge
   1 ILLUMINA (Illumina NovaSeq 6000) run: 37.8M spots, 11.4G bases, 3.6Gb downloads
   Accession: SRX24151665
5. seed sludge
   1 ILLUMINA (Illumina NovaSeq 6000) run: 35.7M spots, 10.8G bases, 3.5Gb downloads
   Accession: SRX24151664
6. Microbial metagenomes of sludge on day 1 of **bioreactor**
   1 ILLUMINA (Illumina HiSeq 2500) run: 20.5M spots, 6.1G bases, 1.8Gb downloads
   Accession: SRX24128112
7. Microbial metagenomes of sludge on day 130 of **bioreactor**
   1 ILLUMINA (Illumina HiSeq 2500) run: 22.4M spots, 6.7G bases, 2Gb downloads
   Accession: SRX24128111
8. Microbial metagenomes of sludge on day 183 of **bioreactor**
   1 ILLUMINA (Illumina HiSeq 2500) run: 23.2M spots, 6.9G bases, 2Gb downloads
   Accession: SRX24128110

Results by taxon

Top Organisms [Tree]
- bioreactor metagenome *(12027)*
- metagenome *(3238)*
- anaerobic digester metagenome *(2919)*
- activated sludge metagenome *(2863)*
- bioreactor sludge metagenome *(2592)*
- All other taxa *(11630)*

More...

Search in related databases

Database	Access		all
	public	controlled	
BioSample	32,597		32,597
BioProject	1,558		1,558
dbGaP			
GEO Datasets	2,403		2,403

Find related data

Database: Select

Find items

Search details

```
Bioreactor[All Fields]
```

Search | See more...

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA | SRA | Search
Advanced | Help

Full | Send to:

SRX24151664: seed sludge
1 ILLUMINA (Illumina NovaSeq 6000) run: 35.7M spots, 10.8G bases, 3.5Gb downloads

Design: seed sludge
Submitted by: Zhejiang Normal University
Study: bioreactor sludge metagenome Metagenome
PRJNA1095645 • SRP499660 • All experiments • All runs
show Abstract
Sample:
SAMN40731501 • SRS20933126 • All experiments • All runs
Organism: sludge metagenome
Library:
Name: SS
Instrument: Illumina NovaSeq 6000
Strategy: OTHER
Source: METAGENOMIC
Selection: other
Layout: PAIRED
Runs: 1 run, 35.7M spots, 10.8G bases, 3.5Gb

Run	# of Spots	# of Bases	Size	Published
SRR28552071	35,685,846	10.8G	3.5Gb	2024-04-04

ID: 32462412

Related information
- BioProject
- BioSample
- Taxonomy

Recent activity
Turn Off | Clear
- thermophli Bioreactor (35269) SRA
- Bioreactor (35269) SRA
- bioreactor (35269) SRA
- bioreactor AND ("platform illumina"[Properties]) (30286) SRA
- thermophilic bioreactor AND ("platform illumina"[Properties]) (462) SRA

See more...

seed sludge (SRR28552071)

Metadata | Analysis | Reads | Data access | FASTA/FASTQ download

Run

Run	Spots	Bases	Size	GC Content	Published	Access Type
SRR28552071	35.7M	10.8G	3.5GB	46.6%	2024-04-04	public

Quality graph (bigger)

This run has 2 reads per spot:

L=151, 100%	L=151, 100%

Legend

Experiment

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX24151664	SS	Illumina	OTHER	METAGENOMIC	other	PAIRED	BLAST

Design:
seed sludge

Biosample

Biosample	Sample Description	Organism	Links
SAMN40731501 (SRS20933126)		sludge metagenome	bioreactor sludge metagenome Metagenome

Bioproject

Bioproject	SRA Study	Title
PRJNA1095645	SRP499660	bioreactor sludge metagenome Metagenome

Abstract:

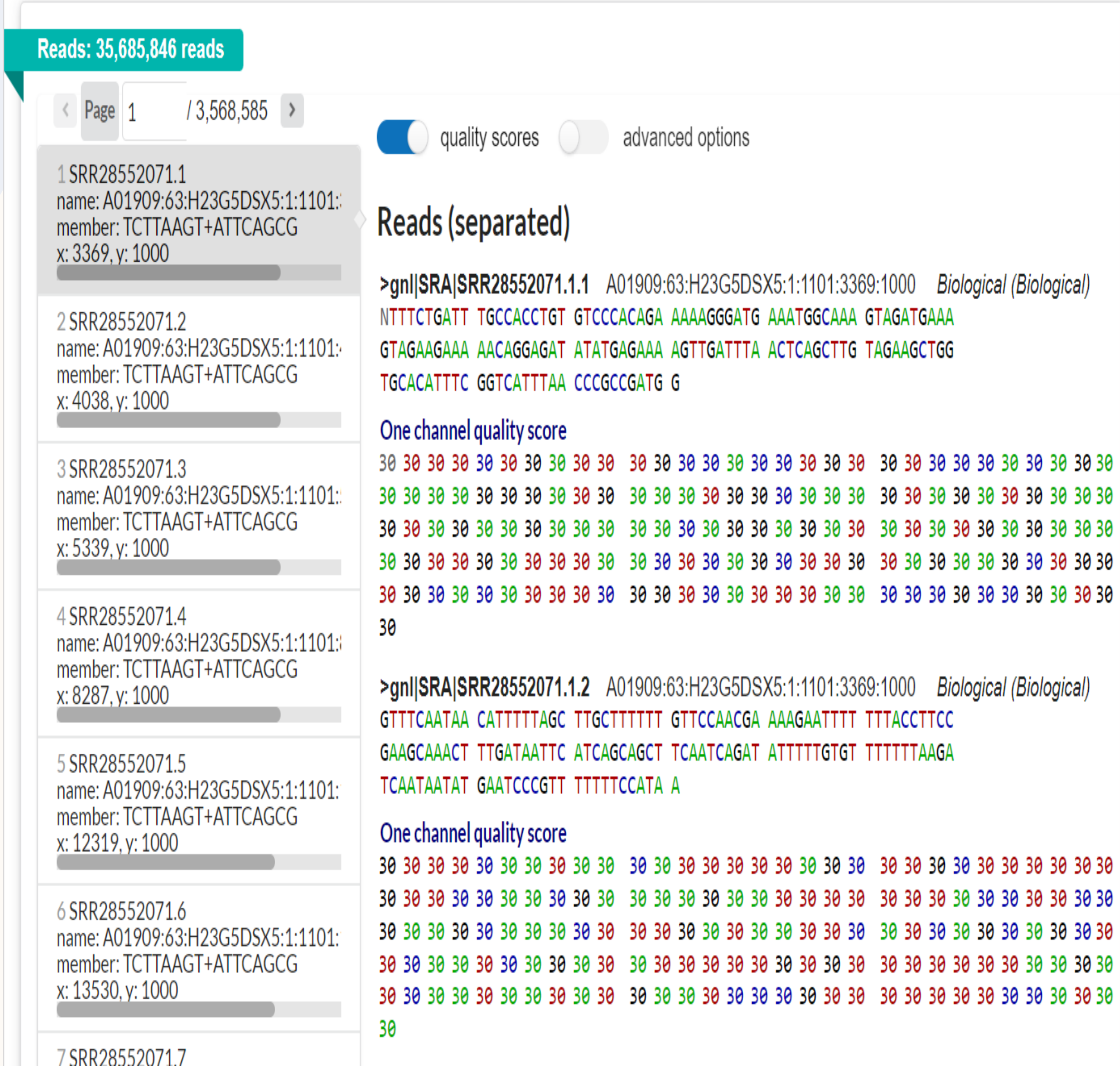
Anammox bioreactor

Taxonomy Analysis

67.54%
IDENTIFIED READS

32.46%
UNIDENTIFIED READS

- cellular organisms: **67.54%**
  - Bacteria: **67.48%**
  - Eukaryota: **0.01%**
  - Archaea: **<0.01%**
- Terrabacteria group: **0.15%**
- FCB group: **0.09%**
- Aeromonas enteropelogenes: **<0.01%** (2 Kbp)
- Oceanisphaera: **<0.01%**
- Nevskiales: **<0.01%**
- Gammaproteobacteria incertae sedis: **<0.01%**
- Moraxellales: **<0.01%**
- Thiotrichales: **<0.01%**
- Legionellales: **<0.01%**
- Acidiferrobacterales: **<0.01%**
- Vibrionales: **<0.01%**
- Cellvibrionales: **<0.01%**
- Salinisphaerales: **<0.01%**
- Pasteurellales: **<0.01%**
- Cardiobacteriales: **<0.01%**
- Hydrogenophilia: **<0.01%**
- Acidithiobacillia: **<0.01%**
- Viruses: **<0.01%**



Reads: 35,685,846 reads

1 SRR28552071.1
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 3369, y: 1000

2 SRR28552071.2
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 4038, y: 1000

3 SRR28552071.3
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 5339, y: 1000

4 SRR28552071.4
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 8287, y: 1000

5 SRR28552071.5
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 12319, y: 1000

6 SRR28552071.6
name: A01909:63:H23G5DSX5:1:1101:
member: TCTTAAGT+ATTCAGCG
x: 13530, y: 1000

7 SRR28552071.7

Reads (separated)

```
>gnl|SRA|SRR28552071.1.1   A01909:63:H23G5DSX5:1:1101:3369:1000   Biological (Biological)
NTTTCTGATT TGCCACCTGT GTCCCACAGA AAAAGGGATG AAATGGCAAA GTAGATGAAA
GTAGAAGAAA AACAGGAGAT ATATGAGAAA AGTTGATTTA ACTCAGCTTG TAGAAGCTGG
TGCACATTTC GGTCATTTAA CCCGCCGATG G
```

One channel quality score

```
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30
```

```
>gnl|SRA|SRR28552071.1.2   A01909:63:H23G5DSX5:1:1101:3369:1000   Biological (Biological)
GTTTCAATAA CATTTTTAGC TTGCTTTTTT GTTCCAACGA AAAGAATTTT TTTACCTTCC
GAAGCAAACT TTGATAATTC ATCAGCAGCT TCAATCAGAT ATTTTTGTGT TTTTTTAAGA
TCAATAATAT GAATCCCGTT TTTTTCCATA A
```

One channel quality score

```
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30  30 30 30 30 30 30 30 30 30 30
30
```

EMBL-EBI home | Services | Research | Training | About us | EMBL-EBI

ENA
European Nucleotide Archive

tp53 Search

Examples: histone, BN000065

Enter accession View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Searching ENA

ENA data can be searched and retrieved interactively and programmatically and visualized using the ENA Browser. Please refer to the following sections for more information about the ENA data access functionality with links to more detailed documentation.

Search term: tp53 Search

Uses EBI Search to perform a free text search across ENA data.

Free Text Search | Advanced Search | Cross References | Sequence Similarity Search | Sequence Version Archive

Search results for tp53

- **Assembly**
  - Assembly (2)
- **Sequence**
  - Sequence (2,686)
  - Sequence (Standard) (2,686)
- **Coding**
  - Coding (3,664)
  - Coding (CON) (363)
  - Coding (Standard) (2,377)
  - Coding (WGS) (540)
  - Coding (TSA) (384)
- **Non-coding**
  - Non-coding (65)
- **Read**
  - Experiment (3,813)
  - Run (2,006)
- **Analysis**
  - Analysis (1)
- **Study**
  - Study (658)
  - Project (1,100)
- **Sample**
  - Sample (1,231)
- **Submission**
  - Submission (Read/Analysis) (1)
- **About**
  - ENA (2)

Assembly View all 2 results.

GCA_014332765.1	ASM1433276v1 assembly for Elephas maximus

Sequence View all 2,686 results.

D83535	Rattus norvegicus Tp53 gene,intron 6.

Sequence (Standard) View all 2,686 results.

D83535	Rattus norvegicus Tp53 gene,intron 6.

Coding View all 3,664 results.

KAI6048876	Marmota monax (woodchuck) TP53

Coding (CON) View all 363 results.

KAI6048876	Marmota monax (woodchuck) TP53

Coding (Standard) View all 2,377 results.

AKI70249	synthetic construct partial TP53

Coding (WGS) View all 540 results.

PRD29181	Nephila clavipes tp53

Coding (TSA) View all 384 results.

MDQ4152535	Cerrophidion petlalcalensis TP53

Non-coding View all 65 results.

CP140814.1:2272184..2272258:tRNA	Rhizobium ruizarguesonis tRNA-Glu

Experiment View all 3,813 results.

SRX19399944	NextSeq 550 sequencing: TP53-MPRA of TP53-KO MCF7 cells

Run View all 2,006 results.

SRR684066	Illumina Genome Analyzer II sequencing: TP53 pro-siRNA

Analysis

ERZ23510811	Single Nucleotide Variants in TP53 gene

Study View all 658 results.

ERP143927	TP53 in lung

Project View all 1,100 results.

PRJNA341515	TP53 RNA-Seq

Coding: PRD29181.1

Nephila clavipes tp53

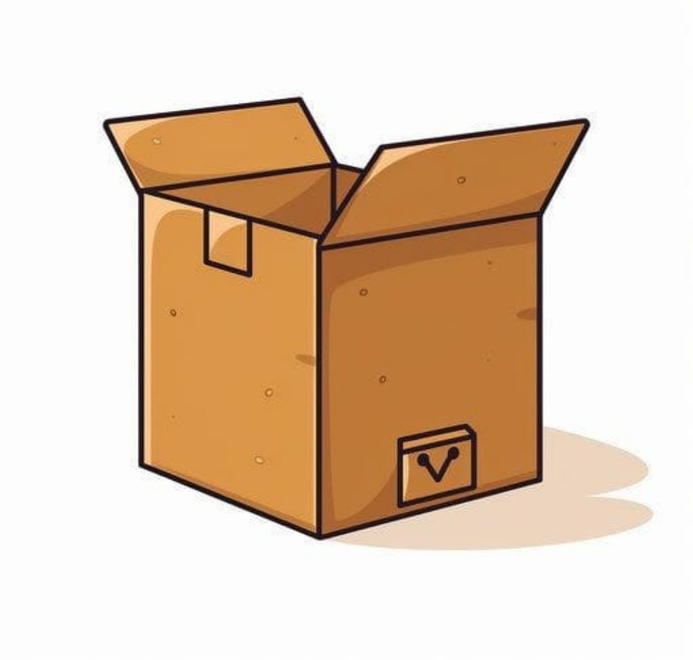
View:	EMBL FASTA
Download:	EMBL FASTA
Navigation:	Show
Publications:	Show

Mol Type:	genomic DNA
Topology:	linear
Base Count:	735
Dataclass:	WGS
Accession:	PRD29181
Country:	USA:Charleston County, South Carolina
Collection Date:	01-Oct-2012
Codon Start:	1
Product:	tp53
Keywords:	WGS
Inference:	ab initio prediction:Maker2:2.31.3 ab initio prediction:SNAP:2013.11.29 alignment:exonerate:2.2.0 alignment:Maker2:2.31.3
Sex:	female
Md5 Checksum:	d168246419033b08ee1973382e3df34a
Collected By:	Linden Higgins
Locus Tag:	NCL1_30185
Dev Stage:	mature
Isolate:	Nep-004
Protein Id:	PRD29181.1
Experiment:	EXISTENCE:RNA sequencing
Tax Division:	INV
Tissue Type:	whole body

Show Less

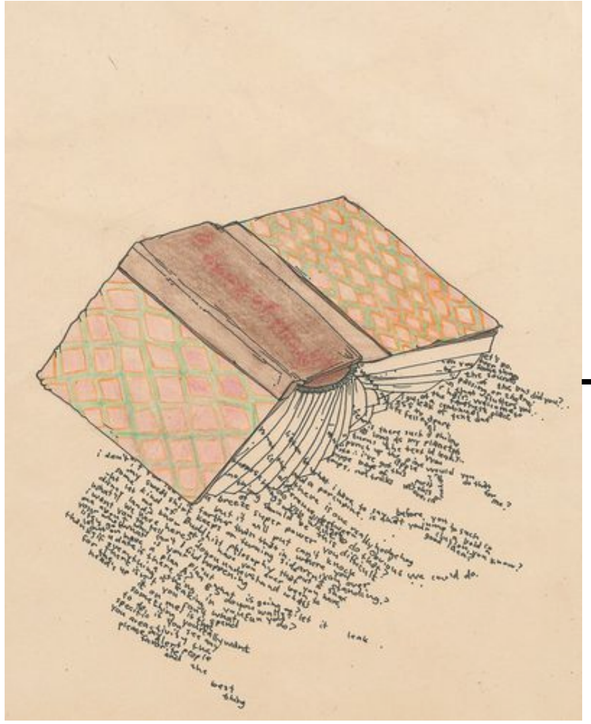
```
ID   PRD29181; SV 1; linear; genomic DNA; WGS; INV; 735 BP.
XX
PA   MWRG01006103.1
XX
PR   Project:PRJNA356433;
XX
DT   15-MAR-2018 (Rel. 136, Created)
DT   15-MAR-2018 (Rel. 136, Last updated, Version 1)
XX
DE   Nephila clavipes tp53
XX
KW   WGS.
XX
OS   Nephila clavipes
OC   Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Chelicerata; Arachnida; Araneae;
OC   Araneomorphae; Entelegynae; Araneoidea; Araneidae; Nephila.
XX
RN   [1]
RA   Babb P.L., Lahens N.F., Correa-Garhwal S.M., Nicholson D.N., Kim E.J.,
RA   Hogenesch J.B., Kuntner M., Higgins L., Hayashi C.Y., Agnarsson I.,
RA   Voight B.F.;
RT   "The Nephila clavipes genome highlights the diversity of spider silk genes
RT   and their complex expression";
RL   Unpublished.
XX
RN   [2]
RA   Babb P.L., Lahens N.F., Correa-Garhwal S.M., Nicholson D.N., Kim E.J.,
RA   Hogenesch J.B., Kuntner M., Higgins L., Hayashi C.Y., Agnarsson I.,
RA   Voight B.F.;
RT   ;
RL   Submitted (02-MAR-2017) to the INSDC.
RL   School of Medicine, Dept. of Sys. Pharmacology and Dept. of Genetics,
RL   University of Pennsylvania, 3400 Civic Center Blvd, Building 421 (SCTR),
RL   Philadelphia, PA 19104, USA
XX
RN   [3]
RA   Babb P.L., Lahens N.F., Correa-Garhwal S.M., Nicholson D.N., Kim E.J.,
RA   Hogenesch J.B., Kuntner M., Higgins L., Hayashi C.Y., Agnarsson I.,
RA   Voight B.F.;
RT   ;
RL   Submitted (02-MAR-2018) to the INSDC.
RL   School of Medicine, Dept. of Sys. Pharmacology and Dept. of Genetics,
RL   University of Pennsylvania, 3400 Civic Center Blvd, Building 421 (SCTR),
RL   Philadelphia, PA 19104, USA
XX
DR   MD5; d168246419033b08ee1973382e3df34a.
DR   BioSample; SAMN06132062.
FT                   /dev_stage="mature"
FT                   /tissue_type="whole body"
FT                   /db_xref="taxon:6915"
FT   CDS             join(MWRG01006103.1:1682..1786,MWRG01006103.1:8814..8997,
FT                   MWRG01006103.1:14089..14199,MWRG01006103.1:18038..18096,
FT                   MWRG01006103.1:19735..19755,MWRG01006103.1:20908..20943,
FT                   MWRG01006103.1:21227..21249,MWRG01006103.1:22319..22378,
FT                   MWRG01006103.1:25449..25573,MWRG01006103.1:25679..25689)
FT                   /codon_start=1
FT                   /locus_tag="NCL1_30185"
FT                   /product="tp53"
FT                   /note="Alias:maker-scaffold_6108-snap-0.35-mRNA-1;
FT                   Maker2_AED:0.39,
FT                   Maker2_QI:309|0.22|0.2|0.3|0.44|0.7|10|0|244; Similar to
FT                   tp53: Cellular tumor antigen p53 (Barbus barbus);
FT                   five_prime_UTR:1373-1681(+)"
FT                   /db_xref="GOA:A0A2P6KS61"
FT                   /db_xref="InterPro:IPR002117"
FT                   /db_xref="InterPro:IPR008967"
FT                   /db_xref="InterPro:IPR011615"
FT                   /db_xref="InterPro:IPR012346"
FT                   /db_xref="UniProtKB/TrEMBL:A0A2P6KS61"
FT                   /inference="ab initio prediction:Maker2:2.31.3"
FT                   /inference="ab initio prediction:SNAP:2013.11.29"
FT                   /inference="alignment:exonerate:2.2.0"
FT                   /inference="alignment:Maker2:2.31.3"
FT                   /experiment="EXISTENCE:RNA sequencing"
FT                   /protein_id="PRD29181.1"
FT                   /translation="MFSYFFRSNWPGEYSFKVSFENQEKNNISKHINWTYSETSNKLYV
FT                   MKDASCPVNFSTNRAMHHGCTVRVMAVYSAPEHFAQPVTRCLNHSRSELEKDVFEAEHL
FT                   IRSESSFASYQTDSGSGRHSVIVPFENPPECITKQGYYTTDEWPDLGLNSQQPPPTKTF
FT                   GHKHVDPNRLKAVSYDSSCAGGPNRRLLTLIFTLELGDIVLGRQSLELKICANPRRDRE
FT                   IAEKRKPEPSASKFKPPEEIII"
XX
SQ   Sequence 735 BP; 240 A; 139 C; 141 G; 215 T; 0 other;
     atgttctcct atttttcag atctaattgg cctggtgaat attcttttaa agtatcattt      60
     gaaaatcaag agaaaaataa tataagcaag cacattaatt ggacttattc agagacttca     120
     aacaaattat atgttatgaa agatgcttct tgccctgtta atttttccac taatagagca     180
     atgcatcatg gttgtactgt gcgagtgatg gctgtgtact ctgccccaga acactttgct     240
     caacctgtga cacgatgtct caatcactca aggtcagaat tagagaaaga tgtgtttgaa     300
     gctgaacatc tcattcgaag tgaaagtagc tttgcatcat atcaaactga ttctgggtct     360
     ggaaggcaca gtgttattgt accattcgaa aatccaccag agtgtattac aaagcaagga     420
     tactacacta cagacgaatg gccagatctt ggcttgaact cccagcaacc tcctcctaca     480
     aaaacatttg gtcacaaaca tgttgatcca aatcgtctca aagcagtatc atatgatagt     540
     tcttgtgctg gtggtccaaa tcgtagactt ttgactctta tattcacctt agaactaggg     600
     gatatagttt tgggaaggca gtcattggaa ttaaagatct gtgcaaatcc aagacgtgat     660
     agagagattg cagaaaagag aaagccagag ccatctgctt caaagttcaa acctcctgaa     720
     gaaataataa tttaa                                                      735
```

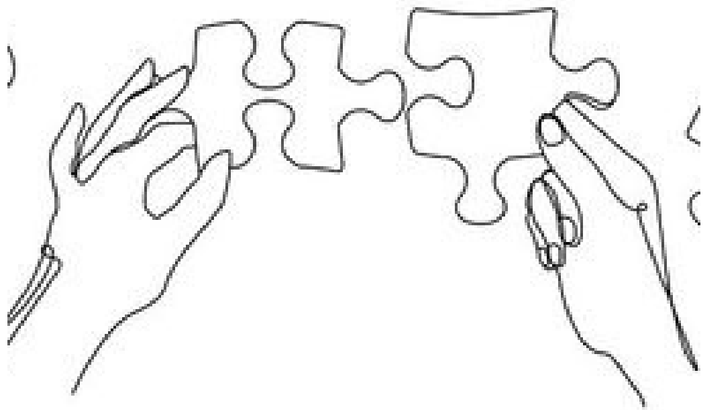
SRA TOOLKIT



"Unlocking the Secrets of DNA, One sequence at a Time!"

Understand SRA toolkit





Need for toolkit

1. SRA $\longrightarrow$ data storage, sharing, and access.
2. Toolkit designed to facilitate the retrieval, processing, and analysis of sequencing in SRA.
3. Tasks $\longrightarrow$ data retrieval, format conversion, quality control, and preprocessing of sequencing data.
4. Toolkit specifically tailored for working with data from the SRA

Lets get started

1. Fetch the tar file from the canonical location at NCBI:

2. Extract the contents of the tar file:

3. append the path to the binaries to your PATH environment variable:

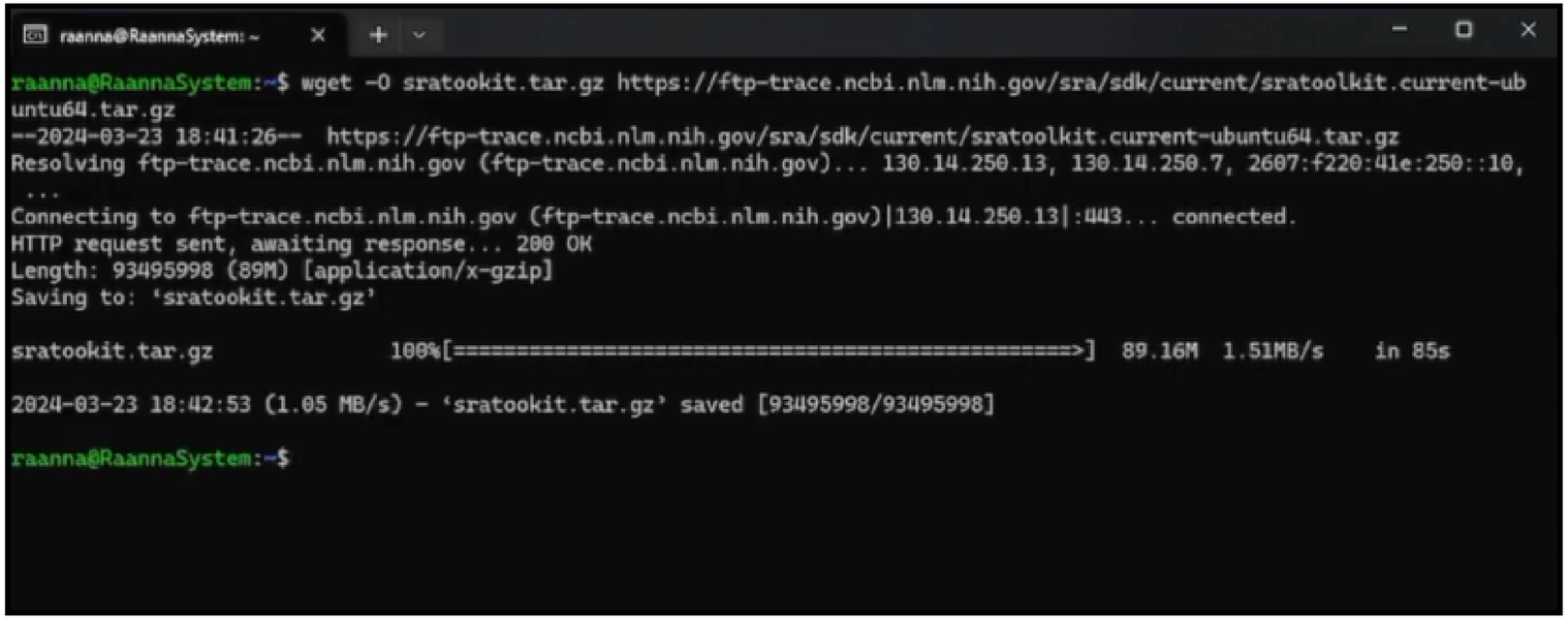
4. Verify that the binaries will be found by the shell:

5 Test that the toolkit is functional

1. Fetch the tar file from the canonical location at NCBI:

Command:

wget --output-document sratoolkit.tar.gz https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz



2. Extract the contents of the tar file:

Command:

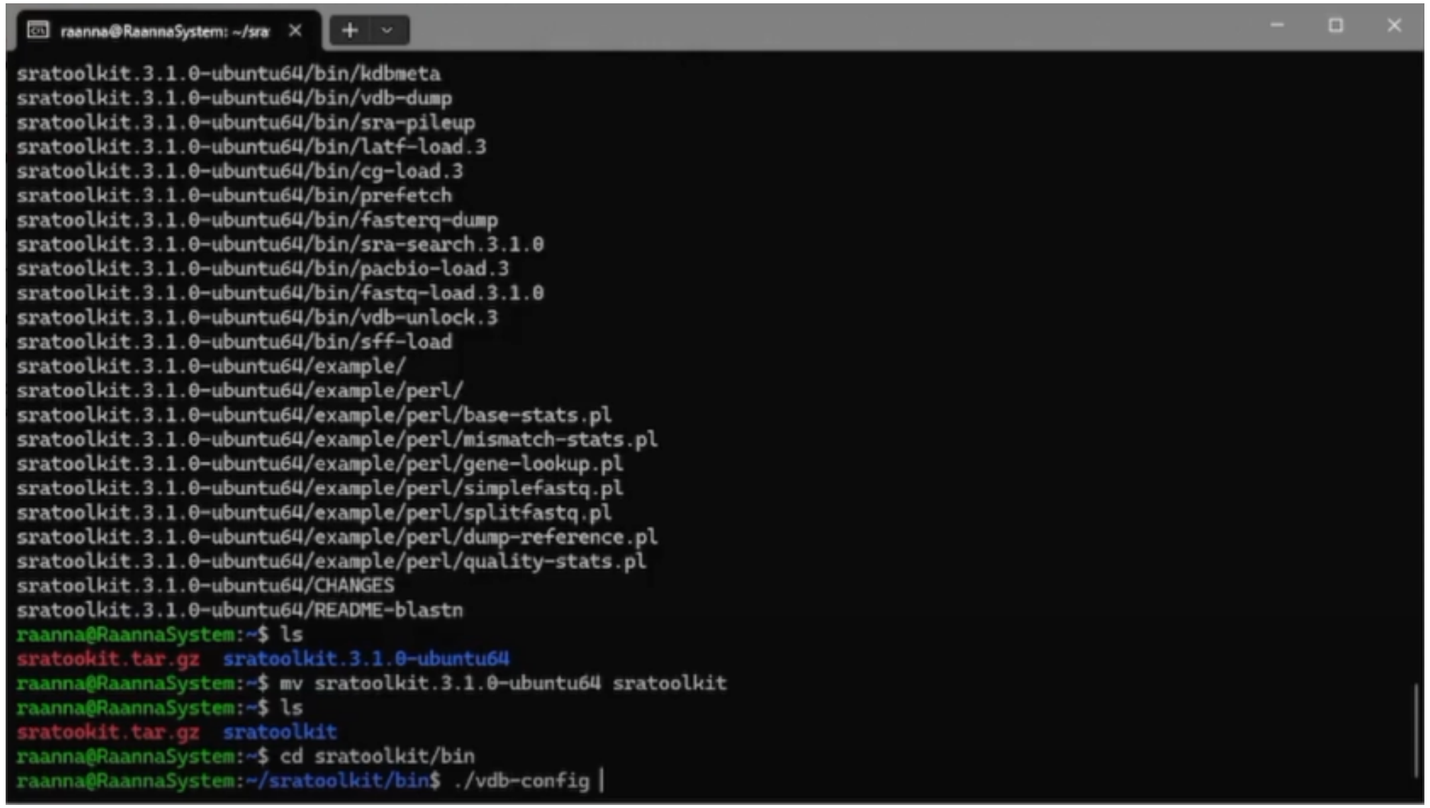
tar -vxzf sratoolkit.tar.gz

```
raanna@RaannaSystem: ~
sratoolkit.3.1.0-ubuntu64/bin/kdbmeta.3.1.0
sratoolkit.3.1.0-ubuntu64/bin/abi-load.3.1.0
sratoolkit.3.1.0-ubuntu64/bin/align-info.3
sratoolkit.3.1.0-ubuntu64/bin/sra-sort
sratoolkit.3.1.0-ubuntu64/bin/kdbmeta
sratoolkit.3.1.0-ubuntu64/bin/vdb-dump
sratoolkit.3.1.0-ubuntu64/bin/sra-pileup
sratoolkit.3.1.0-ubuntu64/bin/latf-load.3
sratoolkit.3.1.0-ubuntu64/bin/cg-load.3
sratoolkit.3.1.0-ubuntu64/bin/prefetch
sratoolkit.3.1.0-ubuntu64/bin/fasterq-dump
sratoolkit.3.1.0-ubuntu64/bin/sra-search.3.1.0
sratoolkit.3.1.0-ubuntu64/bin/pacbio-load.3
sratoolkit.3.1.0-ubuntu64/bin/fastq-load.3.1.0
sratoolkit.3.1.0-ubuntu64/bin/vdb-unlock.3
sratoolkit.3.1.0-ubuntu64/bin/sff-load
sratoolkit.3.1.0-ubuntu64/example/
sratoolkit.3.1.0-ubuntu64/example/perl/
sratoolkit.3.1.0-ubuntu64/example/perl/base-stats.pl
sratoolkit.3.1.0-ubuntu64/example/perl/mismatch-stats.pl
sratoolkit.3.1.0-ubuntu64/example/perl/gene-lookup.pl
sratoolkit.3.1.0-ubuntu64/example/perl/simplefastq.pl
sratoolkit.3.1.0-ubuntu64/example/perl/splitfastq.pl
sratoolkit.3.1.0-ubuntu64/example/perl/dump-reference.pl
sratoolkit.3.1.0-ubuntu64/example/perl/quality-stats.pl
sratoolkit.3.1.0-ubuntu64/CHANGES
sratoolkit.3.1.0-ubuntu64/README-blastn
raanna@RaannaSystem:~$ ls
sratookit.tar.gz  sratoolkit.3.1.0-ubuntu64
raanna@RaannaSystem:~$ mv sratoolkit.3.1.0-ubuntu64 sratool
```

Configuration of the toolkit:

Go to the "bin" subdirectory for the Toolkit and run the following command line:

Command: ./vdb-config -i



Use TAB button to go through the options

Go for tools --- current directory, enter save the changes press esc to exit

3.Append the path to the binaries to your PATH environment variable:

Command: export PATH=$PATH:$PWD

kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin

```
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ export PATH=$PATH:$PWD
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ which fastq-dump
/home/kaveri/sratoolkit/bin/fastq-dump
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ fastq-dump --stdout -X 2 SRR390728
Read 2 spots for SRR390728
Written 2 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
```

4. Verify that the binaries will be found by the shell:

Command: which fastq-dump

Output: /home/kaveri/sratoolkit/bin/fastq-dump

kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin

```
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ export PATH=$PATH:$PWD
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ which fastq-dump
/home/kaveri/sratoolkit/bin/fastq-dump
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ fastq-dump --stdout -X 2 SRR390728
Read 2 spots for SRR390728
Written 2 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
```

Now time to test toolkit!

Connected to database ---- retrieval and analaysing

Command: fastq-dump --stdout -X 2 SRR390728

kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin

```
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ fastq-dump --stdout -X 2 SRR390728
Read 2 spots for SRR390728
Written 2 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ prefetch SRR390728
```

let us try with a couple of tools

1.How can I download the SRA file directly from cmd Linux???

Command: $prefetch SRR390728

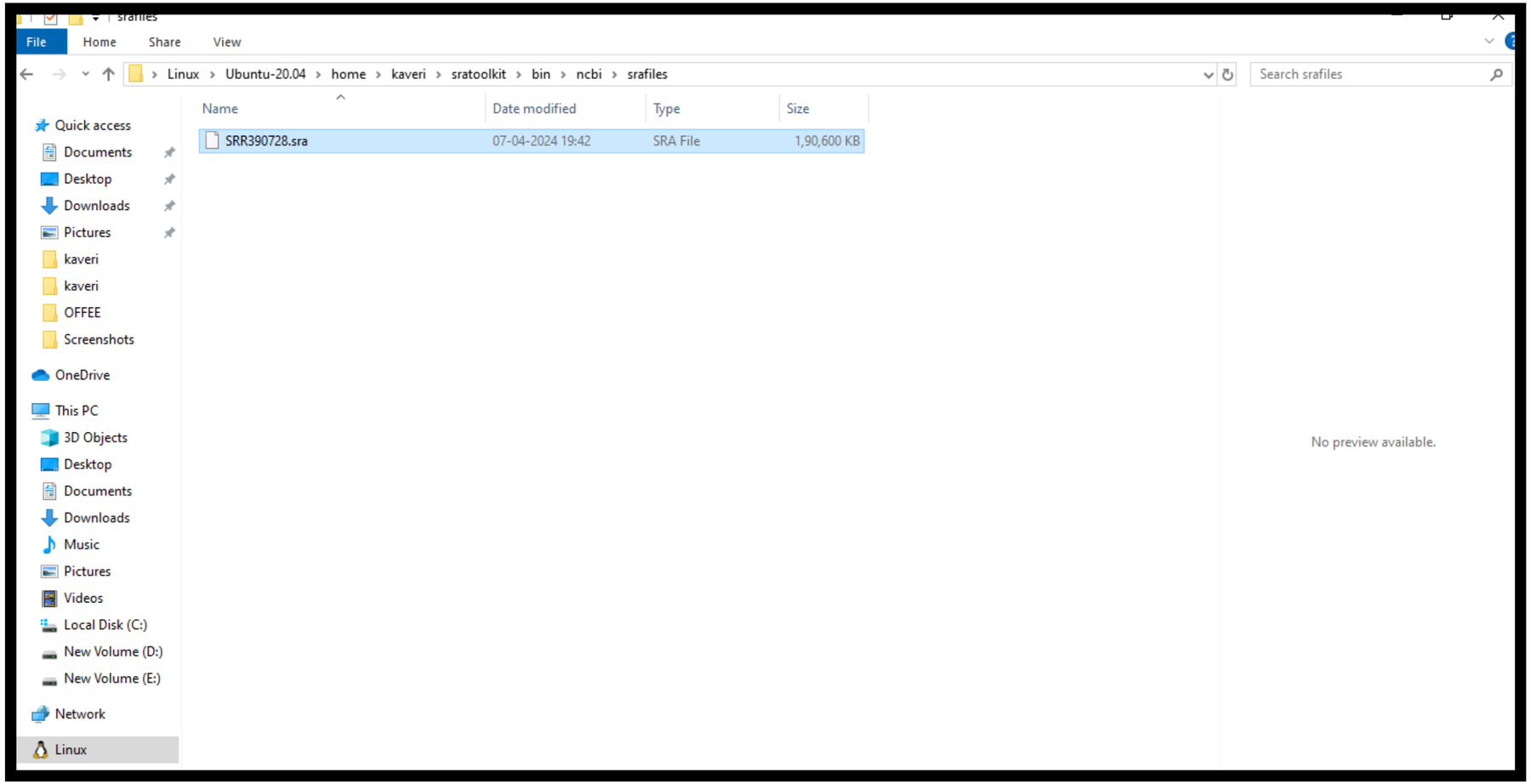
```
kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ prefetch SRR390728

2024-04-07T14:04:31 prefetch.3.1.0: Current preference is set to retrieve SRA Normalized Format files with full base quality scores.
2024-04-07T14:04:33 prefetch.3.1.0: 1) Downloading 'SRR390728'...
2024-04-07T14:04:33 prefetch.3.1.0: SRA Normalized Format file is being retrieved, if this is different from your preference, it may be due to curren
t file availability.
2024-04-07T14:04:33 prefetch.3.1.0:  Downloading via HTTPS...
2024-04-07T14:12:36 prefetch.3.1.0:  HTTPS download succeed
2024-04-07T14:12:37 prefetch.3.1.0:  'SRR390728' is valid
2024-04-07T14:12:37 prefetch.3.1.0: 1) 'SRR390728' was downloaded successfully
2024-04-07T14:13:36 prefetch.3.1.0: 'SRR390728' has 25 unresolved dependencies
2024-04-07T14:13:36 prefetch.3.1.0: 2) Downloading 'ncbi-acc:GPC_000000394.1?vdb-ctx=refseq'...
2024-04-07T14:13:36 prefetch.3.1.0:  Downloading via HTTPS...
2024-04-07T14:14:17 prefetch.3.1.0:  HTTPS download succeed
2024-04-07T14:14:17 prefetch.3.1.0: 2) 'ncbi-acc:GPC_000000394.1?vdb-ctx=refseq' was downloaded successfully
2024-04-07T14:14:17 prefetch.3.1.0: 3) Downloading 'ncbi-acc:GPC_000000395.1?vdb-ctx=refseq'...
2024-04-07T14:14:17 prefetch.3.1.0:  Downloading via HTTPS...
2024-04-07T14:14:26 prefetch.3.1.0:  HTTPS download succeed
```

This will create SRA file or you choose to download where you want

You will have a file in File explorer as type SRA

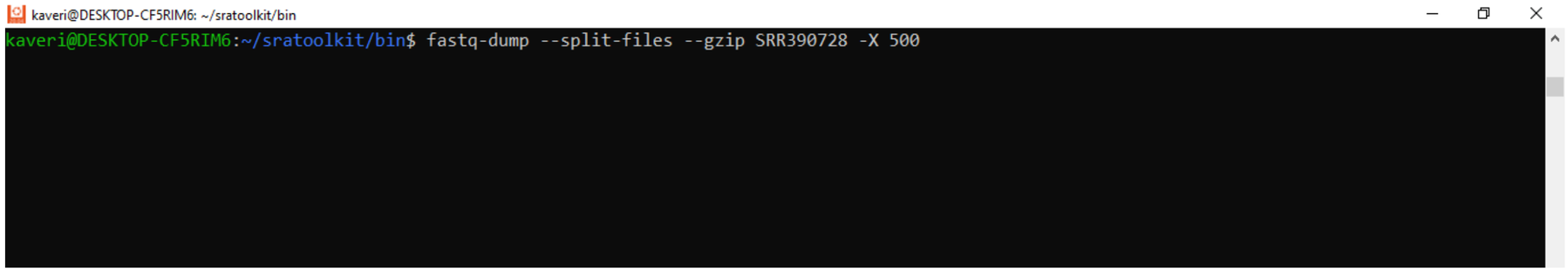
Output:


srafiles
File
Home
Share
View
Linux › Ubuntu-20.04 › home › kaveri › sratoolkit › bin › ncbi › srafiles
Search srafiles
Quick access
Documents
Desktop
Downloads
Pictures
kaveri
kaveri
OFFEE
Screenshots
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
New Volume (D:)
New Volume (E:)
Network
Linux
Name
Date modified
Type
Size
SRR390728.sra
07-04-2024 19:42
SRA File
1,90,600 KB
No preview available.

2.Lets try with downloading Fastq file

This assumes that you have previously 'prefetched' the accession into the current working directory.

Command: fasterq-dump SRR390728



Output:

```
kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ fastq-dump --split-files --gzip SRR390728 -X 500
Read 500 spots for SRR390728
Written 500 spots for SRR390728
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ ls
SRR1951777            dump-ref-fasta                    illumina-load        rcexplain.3.1.0        sra-sort.3          vdb-copy.3
SRR390728             dump-ref-fasta.3                  illumina-load.3      sam-dump               sra-sort.3.1.0      vdb-copy.3.1.0
SRR390728_1.fastq.gz  dump-ref-fasta.3.1.0              illumina-load.3.1.0  sam-dump-orig.3.1.0    sra-stat            vdb-decrypt
SRR390728_2.fastq.gz  fasterq-dump                      kar                  sam-dump.3             sra-stat.3          vdb-decrypt.3
abi-dump              fasterq-dump-orig.3.1.0           kar.3                sam-dump.3.1.0         sra-stat.3.1.0      vdb-decrypt.3.1.0
abi-dump.3            fasterq-dump.3                    kar.3.1.0            sff-dump               srapath             vdb-dump
abi-dump.3.1.0        fasterq-dump.3.1.0                kdbmeta              sff-dump.3             srapath-orig.3.1.0  vdb-dump-orig.3.1.0
abi-load              fasterq.tmp.DESKTOP-CF5RIM6.6603  kdbmeta.3            sff-dump.3.1.0         srapath.3           vdb-dump.3
abi-load.3            fasterq.tmp.DESKTOP-CF5RIM6.6724  kdbmeta.3.1.0        sff-load               srapath.3.1.0       vdb-dump.3.1.0
abi-load.3.1.0        fastq-dump                        latf-load            sff-load.3             sratools            vdb-encrypt
align-info            fastq-dump-orig.3.1.0             latf-load.3          sff-load.3.1.0         sratools.3          vdb-encrypt.3
align-info.3          fastq-dump.3                      latf-load.3.1.0      sra-pileup             sratools.3.1.0      vdb-encrypt.3.1.0
align-info.3.1.0      fastq-dump.3.1.0                  ncbi                 sra-pileup-orig.3.1.0  srf-load            vdb-lock
bam-load              fastq-load                        pacbio-load          sra-pileup.3           srf-load.3          vdb-lock.3
bam-load.3            fastq-load.3                      pacbio-load.3        sra-pileup.3.1.0       srf-load.3.1.0      vdb-lock.3.1.0
bam-load.3.1.0        fastq-load.3.1.0                  pacbio-load.3.1.0    sra-search             test-sra            vdb-unlock
cache-mgr             helicos-load                      prefetch             sra-search.3           test-sra.3          vdb-unlock.3
cache-mgr.3           helicos-load.3                    prefetch-orig.3.1.0  sra-search.3.1.0       test-sra.3.1.0      vdb-unlock.3.1.0
cache-mgr.3.1.0       helicos-load.3.1.0                prefetch.3           sra-sort               vdb-config          vdb-validate
cg-load               illumina-dump                     prefetch.3.1.0       sra-sort-cg            vdb-config.3        vdb-validate.3
cg-load.3             illumina-dump.3                   rcexplain            sra-sort-cg.3          vdb-config.3.1.0    vdb-validate.3.1.0
cg-load.3.1.0         illumina-dump.3.1.0               rcexplain.3          sra-sort-cg.3.1.0      vdb-copy
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$
```

3.To access the SRA query:

Command: fastq-dump --stdout -X 2 SRR390728

```
kaveri@DESKTOP-CF5RIM6: ~/sratoolkit/bin
fasterq-dump quit with error code 3
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$ fastq-dump --stdout -X 2 SRR390728
Read 2 spots for SRR390728
Written 2 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
kaveri@DESKTOP-CF5RIM6:~/sratoolkit/bin$
```

Aspera connect

- Software application for high-speed data transfer over the internet.
- It is not directly related to the SRA Toolkit but can be used alongside it for faster downloads of SRA files from NCBI servers.
- Uses proprietary FASP technology for accelerated data transfer, especially over long distances and high-latency networks.
- It offers features like automatic retry and resume capabilities for enhanced reliability.
- Prioritizes security with robust encryption mechanisms to protect data during transit.

- It provides a user-friendly interface and integrates seamlessly with web browsers.
- Compatible with multiple operating systems, including Windows, macOS, and Linux.
- Integrating Aspera Connect with the SRA Toolkit improves efficiency, reliability, and security for downloading SRA files.

THANK YOU!